BS/BSC

in Applied AI and Data Science

Foundations of Statistics & Probability

IIT Jodhpur

Powered by Futurense

AIL1020

Foundations of Statistics & Probability

Instructor

Dr. Rajlaxmi Chouhan

Associate Professor, Department of Electrical Engineering
Head, Center for Education & Technology for Education
IIT Jodhpur

rajlaxmichouhan@iitj.ac.in

Powered by Futurense

Module 02

Measures of Central Tendency

Exploring data variability

Recap

Data organization and visualization

Introduction to Descriptive Statistics

- Mean, median, and mode
- Population vs. Sample Variance

In this video,

Chebychev's Identity

Apply Chebyshev's Inequality to datasets with unknown distributions.

Interpret the probability bounds for real-world scenarios

Standard Deviation & Probability

Powered by Futurense

Standard Deviation & Probability

The notation $P(X > a)$ represents the **probability that the random variable X takes a value greater than a.**

- X is a **random variable**, meaning it represents different possible outcomes of a process (e.g., dice roll, test scores, AI response time).
- $P(X > a)$ means we are calculating the probability that X will be greater than a specific value a.
- It is equivalent to asking:

  > **"What is the chance that X will be more than a?"**

Powered by Futurense

Chebychev's Identity

Chebyshev's Inequality is a fundamental theorem in probability that provides a bound on how much of the data deviates from the mean.

Chebyshev's Inequality states that **for any dataset** (not necessarily normal),

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

X = A random variable (e.g., test scores, stock prices, etc.)

μ = Mean (average value of the dataset)

σ = Standard deviation (spread of data)

k = Number of standard deviations away from the mean

The probability that a value is at least *k* standard deviations away from the mean is **at most 1/k^2**.

Why should you know about **Chebychev's Inequality?**

- It does not assume normality, making it broadly applicable.
- It gives a worst-case bound, meaning it guarantees that extreme values do not occur too often (helps in setting safety bounds)
- It is widely used in anomaly detection, AI model evaluation, and risk assessment.

Powered by Futurense

Example Fraud Detection in AI Systems

A company tracks daily transactions of an e-commerce AI system.

The mean transaction value is:

μ = 200 dollars

σ = 50 dollars

Fraudulent transactions usually have **extremely high** or **low** values.

E.g. How often a transaction will be more than **3 standard deviations** away from the mean?

AI fraud detection models can use this to set risk thresholds.

At most **11.11%** of transactions will be below **50 dollars** or above **350 dollars**.

$$P(|X - 200| \geq 3(50)) \leq \frac{1}{3^2} = \frac{1}{9} \approx 11.11\%$$

Scenario AI-powered Delivery System

Powered by Futurense

Scenario AI-powered Delivery System

An AI-driven delivery system predicts estimated delivery times (in minutes) for a food delivery app.

After analyzing data from **thousands of orders**, you find:

The mean delivery time (μ) is **30 minutes**.

The standard deviation (σ) is **5 minutes**.

Powered by Futurense

Scenario AI-powered Delivery System

The mean delivery time (μ) is **30 minutes**.

The standard deviation (σ) is **5 minutes**.

Your company wants to guarantee customers that deliveries will be within 20 minutes to 40 minutes most of the time.

Use Chebyshev's inequality to determine the **maximum proportion** of deliveries that might fall **outside** this range.

Powered by Futurense

Scenario AI-powered Delivery System

The mean delivery time (μ) is **30 minutes.**

The standard deviation (σ) is **5 minutes.**

Your company wants to guarantee customers that deliveries will be within 20 minutes to 40 minutes most of the time.

Use Chebyshev's inequality to determine the **maximum proportion** of deliveries that might fall **outside** this range.

Powered by Futurense

Scenario AI-powered Delivery System

The mean delivery time (μ) is **30 minutes.**

The standard deviation (σ) is **5 minutes.**

Your company wants to guarantee customers that deliveries will be within 20 minutes to 40 minutes most of the time.

Use Chebyshev's inequality to determine the **maximum proportion** of deliveries that might fall **outside** this range.

- The number of standard deviations away from the mean:

$$k = \frac{40 - 30}{5} = 2$$

- Using Chebyshev's inequality:

$$P(|X - 30| \geq 2\sigma) \leq \frac{1}{2^2} = \frac{1}{4} = 0.25$$

Scenario AI-powered Delivery System

The mean delivery time (μ) is **30 minutes.**

The standard deviation (σ) is **5 minutes.**

Your company wants to guarantee customers that deliveries will be within 20 minutes to 40 minutes most of the time.

Use Chebyshev's inequality to determine the **maximum proportion** of deliveries that might fall **outside** this range.

$$P(|X - 30| \geq 2\sigma) \leq \frac{1}{2^2} = \frac{1}{4} = 0.25$$

Ans: At most **25%** of deliveries might take **less than 20 minutes or more than 40 minutes.**

Scenario AI-powered Delivery System

The mean delivery time (μ) is **30 minutes**.

The standard deviation (σ) is **5 minutes**.

The company considers a delivery **"very late"** if it takes more than **3 standard deviations from the mean**.

Use Chebyshev's inequality to find out: at most, what **percentage** of deliveries will take more than **45 minutes**?

Powered by
Futurense

Scenario AI-powered Delivery System

The mean delivery time (μ) is **30 minutes.**

The standard deviation (σ) is **5 minutes.**

The company considers a delivery **"very late"** if it takes more than **3 standard deviations from the mean**.

Use Chebyshev's inequality to find out:

At most, what **percentage** of deliveries will take more than **45 minutes**?

Scenario AI-powered Delivery System

The mean delivery time (μ) is **30 minutes**.

The standard deviation (σ) is **5 minutes**.

The company considers a delivery **"very late"** if it takes more than **3 standard deviations from the mean**.

Use Chebyshev's inequality to find out:

At most, what **percentage** of deliveries will take more than **45 minutes**?

Using Chebyshev's inequality:

$$P(|X - 30| \geq 3\sigma) \leq \frac{1}{3^2} = \frac{1}{9} \approx 0.1111$$

At most **11.11%** of deliveries might take more than **45 minutes**.

Summary

Chebyshev's Inequality is a powerful tool when data distributions are unknown.

It provides **minimum guaranteed probability bounds** for datasets.

Coming up next...

Normal distributions and paired datasets